Corpus analysis and conceptual relation patterns

Anne Condamines

Equipe de Recherche en Syntaxe et Sémantique,
CNRS et Université Toulouse Le Mirail
Maison de la Recherche, 5 Allées Antonio Machado
F-31058 Toulouse
anne.condamines@univ-tlse2.fr

Abstract: The aim of this paper is to shed light on the notion of "conceptual relation pattern" via corpus analysis experiments. On the basis of this bottom-up approach, three main points are discussed. First the degree of dependency between conceptual relation pattern and corpus is discussed. This dependency may range from insignificant to complete; it may also be linked to corpus genre. Then, the limits of purely binary conception of relations are examined through the description of patterns taking into account argument structure. Finally, an example in which application may influence patterns choice is presented. Some generally admitted classical assumptions are discussed and revisited by presenting empirical results.

Key words: conceptual relation pattern, corpus analysis, corpus/pattern dependency, corpus genre, empirical approach.

1- Introduction

The problem of elaborating relational systems from corpora with a linguistic method poses questions about a three-way dependency existing between corpus, relations and patterns. In this paper, we explore the potential dependency between corpus and pattern; more precisely, conclusions and reflections about patterns from several experiments in building such a relational system from corpora are presented. These bottom-up studies throw a new light on a number of issues concerning mainly the link between corpus and conceptual patterns but also the problem of binarity and of the link between application and patterns.

The notion of “conceptual relation pattern” is used in a number of disciplines: linguistics, terminology, information retrieval, etc. In each case, the aim is to associate lexical structures with a conceptual relation (or "semantic" or "lexical", depending on the discipline) and to use these structures in order to spot these relations. This notion is based on assumptions which -while not always clear to the authors who use it- nevertheless certainly influence their studies. Experiments in corpus analysis for identifying conceptual relations show that knowledge patterns are much more complex than is generally supposed and that many of the assumptions need to be reconsidered. Corpus analysis currently claims to take into account three kinds of elements which affect the study of relations and their patterns. First, a corpus is not simply a set of utterances in a language. It has been created with a precise objective and contains texts written by authors with a well-defined aim, at a precise time and for a particular set of readers. In brief, a corpus is linked together by significant extra-linguistic characteristics (named *genre* by Biber (Biber 1988)). Second, as a set of discursive realizations, a corpus appears in a linear form, as a set of linguistic elements entering into numerous and varied semantico-syntactic relations. Now, the notion of “relation” is generally considered allowed us to build a system with lexical units that do not have syntactic relations since they are no longer in discourse. This transition from discourse to model (or *metalinguistic level* as several authors call it), is the heart of the problem of identifying relations with patterns. But many studies address the issue the other way round. Their point is to see how the transition from the linguistic system to discourse manifestations is achieved using an introspective method. This results in many interesting questions being left unanswered. The third element to take into account when analyzing conceptual relation patterns in the corpus is the aim of the modeling which may vary according to the point of view adapted. Very often, different structures can be considered:

“ Dans la langue tout est mouvance, tout est flou. Et, du fait même les axiomatisations possibles sont en grand nombre – aussi artificielles les unes que les autres, ce qui ne signifie pas qu’elles soient dénuées d’intérêt, car chacune éclaire le réel sous un certain angle » (Martin 1983 : 86)

[In language, all is moving, all is blurred. And so the possible axiomatizations are very numerous, all equally artificial, which does not mean that they are without interest since each of them gives a particular view of reality].

The transition from a textual form to a metalinguistic one is always the result of a particular interpretation, the latter often guided by an interpretational aim.

So, from a bottom-up point of view, the problem of conceptual relation patterns is influenced by three elements which characterize corpus analysis (extra-linguistic characteristics, discursive nature and aim of the analysis). With the help of examples, this paper tries to present how these characteristics play a role in the identification of conceptual relation patterns from corpus.

Section 2 exposes the classical view on conceptual relation patterns and the assumptions it imposes. Section 3 deals with pattern/corpus dependency when it is insignificant or complete; in Section 4, two examples of dependency between patterns and corpus genre are presented. Section 5 poses the problem of a purely binary conception of relations and Section 6 shows how application may influence patterns choice.

2- The notion of "conceptual relation pattern" and the assumptions it imposes.

The notion of "conceptual relation pattern" appears under different names in various studies, some of which are indicated in Section 2.1. In Section 2.2, the resulting assumptions are highlighted.

2-1 The notion of "conceptual relation pattern"

In linguistics, the notion of relation originates from the work of structuralists who consider the language as a system of signs (after Saussure). In fact, even more than signs, relations are fundamental in such a perspective:

« Chacune des unités d'un système se définit ainsi par l'ensemble des relations qu'elle soutient avec les autres unités, et par les oppositions où elle entre » (Benveniste 1966 : 21)

[All system units are defined by a set of relations in contact with others and by the oppositions in which they appear].

Relations operate at each level of the language (phonological, morphological, syntaxical and lexical) since they allow systems to be elaborated.

The notion of "mark" also appears in the context of structuralism, firstly in phonology where it designates the feature allowing the distinction between two phonemes. In morphology, probably by analogy, the element bearing a formal trace distinguishing it from the non-marked one was named a marked element (*lion/lioness, kitchen/kitchenette*). Then the idea of a form appeared possibly with a content and allowing the differentiation of one element from another. By its regular use, this form may be considered as the mark of a distinctive phenomenon from which it is possible to build a system.

At semantico-syntaxic level, the notion of "conceptual relation patterns" may be considered as the equivalent of "marks" at morphological level. They are recurrent forms used to indicate differences between lexical units and to build a structure. In comparison with morphological marks, semantico-syntaxic forms may be various and non-predictable. This notion of "mark" or "pattern", used to identify distinctive elements (in the tradition of structural analysis), can also be used to spot similarities. Within a relational structure, there are not only differences but, above all, there are proximities which justify the relations. At morphological level, similarities themselves are formal (and thereby marked). Marked and unmarked elements have the same root and only the suffix varies. Things are different at the semantico-syntaxic level where similarities are semantic and therefore unmarked. Consequently, the "pattern" -the only formal trail of a semantic relation- is useful for identifying differences as similarities. This fact raises many questions and interpretational difficulties. It would be preferable if the pattern were stable, always referring to the same content. This would allow us to use it in a systematic way, to identify both proximity and difference. Indeed, many studies are based on just this assumption. Unfortunately, patterns do not always work in this way (cf below).

The idea of associating lexical elements with a semantic relation is found under different names in different disciplines: linguistics, terminology, computer science

(within knowledge acquisition from texts). To give a few linguistics examples, Lyons (Lyons 1978) was probably the first to evoke particular "formulae" used for fitting terms linked by a hyperonymic relation. Cruse (Cruse 1986) speaks of "diagnostic frames".

In terminology, the use of particular linguistic patterns to identify relations between terms systematically is rather recent. It appeared when issues within textual terminology and knowledge acquisition from texts were identified as being similar. This was done by creating the concept of a “terminological knowledge base” (Meyer and al. 1992). Authors speak of "knowledge rich contexts" (Davidson and al 1998) or "connective phrases" (Pearson 1998). We referred to "marqueurs de relation conceptuelle" ("conceptual relation patterns") or just “marqueurs” (patterns) (Condamines and Rebeyrolle 2000). In computer science, more specifically in works on knowledge acquisition from texts, expressions such "explicit relation marker" (Bowden and al. 1996) or "knowledge probes" (Ahmad 1992) is used. The labels used show two different points of view, one focusses on discourse, the other emphasizes the transition from discourse to a model (metalinguistic level). In the first case, patterns are considered in terms of discursive structures ("expressions", "formulae"). There is no intention of using these contexts for building an abstract model. In the other case, this idea is very important: "knowledge probes", "diagnostic frames". Whatever the name chosen, all the authors are aware that it is possible to build an abstract model from discursive tracks. The notion of “pattern” is then highlighted as an indication of the possible transition from discourse to meta-discourse. This elaboration supposes a reduction in the number of discursive structures, in order to preserve only a reduced set of relations. Patterns are not only clues to a potential structure with explicit relations, but they are also lead to the attribution of a meaning –that might be more or less obvious- to this relation. The nature of the interpretation is quite clear when the pattern is not very context-dependent (linguistic context, nature of the corpus and the aim of the modeling). On the other hand, the help provided for interpreting is unclear when the pattern is highly context-dependant, as will be demonstrated in the rest of this paper. Finally, for the purposes of corpus analysis, we can define conceptual relation patterns as follows:

Conceptual relation pattern: a discursive structure used as an indication of the possible transition from the discourse to a model, allowing, the more or less direct construction of a model in the form of a semantic relation depending on its relation with the context.

2.2 Assumptions made under the notion of "conceptual relation pattern"

The majority of authors, using a top-down view (from system to its discursive realization), associate the notion of "pattern" with assumptions which do not even cross the author's mind. These assumptions are:

1- There is one linguistic system and one pattern system. Most studies resorting to the notion of "patterns" are based on the idea that there is just one semantic system guaranteeing consistency. One of the main objectives of this article is to show that it is necessary to moderate this point of view when the study is conducted on real corpora.

2- Linguistic relations are predictable by introspection. Viewing language as a unified semantic system may lead to the idea that patterns (as well as relations) may be described in an introspective manner and so be predictable. But, as shown below, in some cases, elements which play the role of patterns are totally unpredictable (see also Meyer 2001).

3- Patterns allow access to explicit relations. The idea here is that the link between form and content is direct. However, there are cases in which a form is well linked to a relation even though this form does not express the sense of this relation.

4- The relations concerned are binary in nature. Indeed, the binary nature of the relation underlies most studies of relations. What is considered interesting is the relations between two units and only between two units. An analysis of the context in which these units appear shows that their binary nature is often insufficient for taking into account the behavior of these units and their patterns into account.

5- Patterns are not linked to the application in mind. Since context is not taken into account in most of works about conceptual relation patterns, the aim of the modeling is not considered. Even at pattern level, this must be taken into account.

The rest of this paper aims to show that these hypotheses must be reexamined if, with a bottom up view, patterns are examined in real corpora in order to describe their

function rather than being first described and then systematically searched within corpora and so considered as general.

3- Most extreme cases: Pattern/corpus dependency is insignificant or complete

Two cases can be used to set limits to the continuum existing within the dependency between the corpus and the patterns used in it: (1) cases where the corpus does not seem to have an influence on patterns and; (2) those where patterns are completely linked with the corpus.

3-1 When the dependency corpus/pattern is weak

Experiments in analyzing corpora within our group as well as results from experiments made by other groups tends to show that some patterns (of some relations) occur in most corpora. Thus, it is possible to deduce (with caution) that there is no dependency between corpus genre and these patterns. This conclusion is also similar to the assumption used *a priori* by authors who are not interested in this potential dependency. But moving from discourse to modeling is not achieved without difficulties.

For example -even if no systematic analysis has been performed using very large corpora- it seems that the definitional context in which hyperonymy appears between *definiens* and *definiendum* plus *differentiae* is explicited (following Aristotle's model), are very frequent within corpora (Rebeyrolle 2000). Following this, it would be appropriate to analyze a corpus starting with this pattern, the most general for hyperonymy:

[N is a N + differences](in the form of an adjective, a relative clause, etc.).

This assertion has to be considered with caution and three comments must be added.

1- The hyperonymy relation is probably the most structuring. Therefore, the high number of occurrences of this relation in corpora is not surprising -even in corpora with no didactic content- simply because definition (which favors this relation), with its clear metalinguistic orientation, may be used each time it is necessary to

clarify a subject and to improve the understanding of a concept. This explains why this relationship is encountered in many corpora, independently of genre.

2- The fact that a pattern (or a relation) is weakly linked to corpus genre does not mean that these patterns will be encountered in all corpora. In 1993, we (Condamines et Amsili 1993) analyzed a corpus from *Matra Marconi Space*. From a set of requirements, written by domain experts for other experts, no known hyperonymic patterns were encountered. It is likely that the expert status meant that it was not necessary to use hyperonymy. However, it seems difficult to predict that this relation will never appear in a corpus written from expert to expert.

3- The presence, in discourse, of elements with the form of a pattern is not sufficient to guarantee that they will be used to build a model. There are at least three reasons for this. First, some patterns are "polysemic", that is, they may refer to several relations. This is the case with the following:

[dét SN1 comme SN2] ([dét NP1 as dét NP2])

indicating a hyperonymy in example (1):

(1) Un département comme la Seine bénéficie à la fois d'arrivées d'enfants et de scolaires (corpus: Atlas scolaire*)*

[A department such as the Seine benefits both from the arrival of children and students].

The same pattern does not indicate the same relation in example (2):

(2) On comprend que les lycées professionnels et d'enseignement général, comme l'université, soient très peu tournés vers les formations scientifiques et technologiques. (corpus: Atlas scolaire).

[We can understand that the professional lycées and general teaching, as well as the University, are little orientated to scientific and technological formations].

Secondly, the aim of modeling guides the interpretating process, which can at the same time result in choosing to retain a context or not and in choosing to model it according to one relation or another.

Thirdly, and this is perhaps the most difficult element to control, it may be difficult to determine if the speaker using conceptual relation patterns is expressing his (or her) own point of view or if he/she assumes the point of view of a group of speakers. Generally, only the second case needs to be modeled because such models must be

acceptable and reusable within collective tasks. However, it is not improper to imagine retaining these same contexts, even if they are used by an isolated speaker, if the objective is to identify the system within the corpus emanating from this speaker (for example, literary works). In this case, interpretation and modeling objectives are crucial.

Finally, apart from hyperonymic patterns in definitional contexts which occur very frequently, other patterns which *a priori* would appear to be common, occur very rarely in some corpora as Séguéla has demonstrated in his Ph.D. (Séguéla 2001).

3-2 When the corpus/pattern dependency is complete

In some corpora, it is proven that very specific structures play the role of pattern for some relations. These structures are obviously unpredictable and only a fine-grained analysis brings them to light: "While certain patterns seem quite logical…others are extremely unpredictable" (Meyer 2001: 295). We found such a structure in a corpus from EDF (Electricity of France), concerning specifications and writing of documents in software. It is a corpus of approximately 350 pages, written by several experts but with the same purpose. The guide is also written with an obvious injunctive style. In this corpus, the pattern for the relation `CONDITION` (reorganized in `conditions-the-beginning-of` and `conditions-the-end-of`) was finally:

[((phase, étape) ou déverbal) + (lorsque, dès que) + V au passif]

[((phasis, stage) or nominalization) + (when, as soon as) + passive V].

Examples:

(3) La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.

[The component integration phase may begin when all the software elements have been written.]

(4) Cette transition est déclenchée par le responsable des développements dès lors que toutes les manifestations spécifiées sont implémentées.

[This transition is triggered by the development manager when all the specified achievements have been implemented.]

Note that this idea of condition is not present in each expression of succession in the corpus:

(5) V est incrémenté lorsque le logiciel subit une évolution ou une modification majeure

[V is incremented when the software is subject to an evolution or major modification].

In such cases, a problem arises from the difficulty of identifying these patterns. Concerning the example above, we explained in Condamines and Rebeyrolle (2000) how we succeeded in achieving this task. To summarize, we applied the following sequence:

- Use of succession patterns (allowing parts of a process to be rearranged (phase)),
- By results analysis, identification of an example clarifying the link between succession and condition:

(6) Conditions de passage à la phase suivante :

La phase d'intégration du produit est achevée lorsque tous les tests prévus au plan d'intégration ont été exécutés avec succès.

[Conditions for passing to the next stage:

The product integration phase has been reached when all the previous tests in the integration plan have been carried out successfully.]

- Finally, description of a "good" pattern, which generates neither noise nor silence.

A certain regularity in modes of expression, potentially seen as conceptual relation patterns, is probably created within groups of speakers, which can either be stable, or constituted for a purpose or for a determined period. These contexts can be identified in corpora only after a minute analysis. In such cases, it is very difficult to anticipate which regularities are going to emerge from a given corpus. It is therefore necessary to develop methods for identifying these regularities rather than proposing a description of patterns *a priori*.

4- When corpus/pattern dependency is linked to corpus genre

Sometimes, regularity in expression is not specific to a particular corpus but rather to a corpus insofar as it is representative of other corpus with identical characteristics. Then

a corpus-based linguistic analysis becomes possible, with the aim of identifying links between the corpus genre and pattern. Two examples of this type will now be presented: the French preposition [chez] and the structure [N1 avec N2].

4-1 The case of [chez]

In some French contexts, *chez* is used to express a meronymic relation:

(7) Chez les colobinés, le nez fait saillie sur la lèvre supérieure.

[In the Colobinés *the nose juts out over the upper lip.]*

In a study presented in (Condamines 2000), a hypothesis was made to the effect that that the co-occurrence *chez/meronymic expression* is just possible in natural sciences corpora. In order to verify this hypothesis, a corpus containing examples with *chez* from texts of different types was built:

- First distinction: natural sciences domain vs ideas and creation domain as in *(8):*

(8) les images des poètes ne roulent pas en torrents ou en avalanche, comme c'est souvent le cas chez Shakespeare

[The images of the poets do not flow in torrents or avalanches, as is often the case in Shakespeare.]

- Second distinction, within the natural sciences domain: didactic vs. non-didactic.

When building the didactic corpus, we took into account all the examples containing *chez* extracted from a number of sections in the *Encyclopaedia Universalis* (1995) natural sciences domain, and all the examples with *chez* extracted from a textbook written by Bernard (1860). After this selection, the set studied comprised 692 examples.

It is clear that the hypothesis concerning a relation between the corpus genre and the presence of a meronymic relation is confirmed. In fact, none of the examples from the "ideas and creation" corpus contains a meronymic relation. Within the natural sciences corpus, just a third of the examples express this relation in the non-didactic corpus and half in the didactic one. A fairly clear correlation appears between the natural sciences corpus with didactic genre, on the one hand and the expression of a meronymic relation, on the other. In other words, anyone interested in meronymy as it is expressed

for living organisms, (*chez* was not used for minerals in the same kind of examples), and has access to a corpus with such characterisitics, should use examples with *chez.* About half of the examples will be directly usable.

Another conclusion was highlighted by this study, leading to a new view of the notion of "pattern". The analysis of real examples shows that meronymic relations in sentences with *chez* is not expressed (asserted) by this relation but, on the contrary, it is considered as known (presupposed). The relation between N1 and N2 in sentences exposing the structure [chez SN1,SN2] is not clarified by the presence of *chez*. In example *(7)*, it is possible to say that there is a meronymic relation between *ne*z and *colobinés*, because the relation is known, not because sentence makes explicit reference to it. Similarly, it is known that there is not a meronymic relation between *préfloraison* and *fraisier* in (9), even though the same structure can be found:

(9) Chez le fraisier, la préfloraison est calvaire pour le calice, quinconciale pour la corolle.

[In *the strawberry plant, the pre-flowering is in* calvary *for the cup, and in staggered rows for the corolla].*

It is thus possible to say that [chez] often indicates a meronymic relation, with conditions about the corpus in which it is used. While the relation is not clarified by [chez], the preposition is not exclusively indicative of a pattern, as we described in Section 2 -that is, marking a possible modeling **and** guiding its interpretation. Nevertheless, it may be used systematically to identify a meronymic relation.

Note that the potential of a pattern to guide interpretation remains an open question. Even in staightforward patterns this is not so clear-cut. See examples (1) and (2) with [comme] (*as*) in 2. Because we know that university is not a kind of lycées, it is possible to know that *comme* does not indicate a hyperonymic relation. That is to say that, in a new domain, it is far from certain that [comme] may be used for extracting hyperonymic relations. In some cases, this possibility is linked with corpus genre, as with [chez]. In others, this possibility may completely aleatory.

4-2 The case of [N1 avec N2] ([N1 with N2])

Even in paradigmatic relations such as hyperonymy and meronymy, considered to be the most frequent, it is possible to discover new patterns during corpora analysis, when the corpora are representative of different genres. This happens with the pattern [N1 avec N2], which is indicative of meronymic relations in a corpus composed of extracts from two toy catalogues. The description of components is obviously important in such a corpus and it is obvious that this pattern is a very efficient way of indicating this kind of meronymy:

(10) Porteur évolutif avec canne, repose-pieds et ceinture de sécurité amovible (la Grande Récré),

[Modulable carrier with cane, footrest and removable safety belt].

(11) Une cuisine moderne avec hotte, plaque de cuisson, coin repas… (Carrefour).

[A modern kitchen with extractor fan, hotplate, breakfast bar, etc.]

Other meronymic patterns, well known, are also used occasionally:

(12) Coffret contenant 3 poupées, 1 jeu de société et de nombreux accessoires

[Box containing three dolls, a card game and numerous accessories.]

(13) DXGalaxy Megazord : composé de 5 zords représentant 5 animaux sauvages.

[DXGalaxy Megazord, containing 5 zords representing 5 wild animals.]

Thus, in a 5,000 word corpus, [N1 avec N2] referring to meronymy is used 63 times, verb [comprendre] 2 times and structure [être composé de] 3 times.

The study is still under way and a variety of aspects have not yet been examined, namely: the accurate description of the pattern, and the identification of corpora genre in which this pattern is relevant.

- Description of the pattern. Two aspects require particular attention:
    - presence or not of a determiner before N2 and nature of the determiner (In 22 out of 63 examples there is a determiner, as in (14):

(14)Poussette d'hôpital avec tous les accessoires de l'infirmière

[Hospital pushchair with all a nurse's accessories.]

    - closeness of N1 to the remainder of the structure. Actually, the paronymic style of these descriptions means that, in some examples, [Avec N2]

appears directly after a period, while N1 is located higher in the sentence or text.

- Identification of corpus characteristics allowing the establishment of a correlation between the pattern and meronymic relation. It is very likely that this pattern remains valid for other corpora, especially in small ads such as "House with garden for sale". But it is difficult to find such corpora in electronic form, due firstly to the difficulty of delimiting relevant characteristics of corpora, and, secondly, to the scarceness of this type of corpora in electronic form.

5- When the relation is not binary

This paragraph deals with cases where a binary character is insufficient to account for the real functioning of relations in corpora. This limit became obvious during a study of a corpus provided by different organizations in charge of traffic flow in Toulouse, France. Our work was to analyze terminological and discursive differences between the various partners. We chose a method allowing us to obtain results rapidly. Based on the examination of the list of nouns provided by the Nomino tool, we came to realize the importance of communication in the corpora of each partner and decided to analyze the following scheme:

`X communicates Y to Z via W`

The tool we used to manage terminological relations was unable to account for such a non-binary system of relations. So, we decided to consider the following relations:

- `X émet Y (X transmits Y)`
- `X médiatise Y (X mediates Y)`
- `X envoie à Y (X transmits to Y)`
- `X reçoit Y (X receives Y).`

Patterns under the traditional form of lexical elements were highlighted for these relations. For example :

- [Y be dét. interlocutor of X] for relation : `X transmits to Y` as in :

  (15)La ville de Toulouse et les ASF sont les interlocuteurs principaux de la DDE (X = DDE, Y = Ville de Toulouse et ASF.

  [The city of Toulouse and the ASFs are the main interlocutors of the DDE (X = DDE, Y = town of Toulouse and ASF)].

- [Y be transfered via X] for relation `X mediates Y`

(16)La nuit et le week-end, les appels sont transférés via des liaisons spécialisées. (X =liaisons spécialisées, Y = appels)

[At night and at weekends, calls are transferred using dedicated links (X = dedicated links, Y = calls].)

Beyond these "traditional" patterns, some other contexts use what could be called semantico-syntactic patterns withmore elaborate structures than simple lexical patterns. For example (17):

(17)Chaque subdivision transmet au CIGT une fiche relative aux chantiers qui se terminent

[Each subdivision transmits to CIGT a form related to completing sites].

may be organised according to two different relations:

- *Subdivision* émet *fiche relative aux chantiers qui se terminent*

(*Subdvision* transmits *a form related to completing sites*)

- *Subdivision* envoie à *CIGT*

(Subdivision transmits to *CIGT).*

Then, the patterns may be the following:

Relation	Pattern
X transmits Y	[X is the subject of a transmitting verb, Y is the object of the same verb and X is [+ human]]
X transmits to Y	[X is subject of a transmitting verb, Y is the indirect object of the same verb and X is [+ human]].

It is also possible to make use of this kind of pattern systematically (and even automatically). Verb role is then fundamental since the associated role of any argument can be used to identify relations between two of them. This kind of pattern requires the construction of lists corresponding to semantic classes. For example, <transmission predicates> or <humans>. Such a requirement is problematic since it relies on semantic knowledge which is not not necessarily stated in the corpus analyzed, and there is a risk that the semantic *a priori* knowledge does not correspond to real linguistic behavior in the corpus. But this problem cannot be dealt with here.

So, it appears clearly that binary relations are restrictive. In fact, in an argumentative scheme, all arguments are interdependent and it would be necessary to take them into account by using a n-ary scheme. These remarks again show that barriers between

disciplines are not impenetrable. Indeed, such recurrent linguistic schemes are used systematically to identify information in a number of applications. Two examples may be given. Firstly, work carried out on man-machine dialogue is based on a semantic grammar approach which uses a predicate argument structure in order to identify information (for example). Another example is work on information retrieval. It does not completely use argument structure (since the analysis is rarely done with a labeled corpus) but predefined lexical structures in which each noun position corresponds to a semantic role (Gaizauskas and Wilks 1998). For naming and defining these structures in this discipline, authors use terms very close to that used to deal with pattern relations in other disciplines. Riloff speaks of "extraction patterns" (Riloff 1996). Basili and Pazienza speak of "surface form", defined as: “Another valid and slightly more comprehensive definition describes the lexicon as an association between surface forms and linguistic information” (Basili and Pazienza 1997,44).
A common factor to all this work on links between discourse and information modeling is that it concerns very specific languages used for specific purposes.

6- When a pattern is linked to the purpose of modeling

Terminologists and especialy computer scientists who are required to model knowledge, know that the choice of one relational structure over another may be linked to the modeling economy or to its purpose. For example, in the senteces quoted above, we decided finally to subdivide the `condition` relation into two sub-relations: `condition the beginning of` and `condition the end of`. An alternative would have been to retain a sigle relation and to specify for each node whether it is the beginning or the end of the given process. This would be much more costly than resorting to two different relations. For example, according to its purpose, the model may be built either to identify inconsistencies (in this case, the modelling must be very precise) or to index other corpora (then, the modeling is more flexible).
What has been much less studied is the fact that the patterns themselves may be differentiated according to the results they give and to their relevance to the modeling aim. Nevertheless, it seems that this dependency exists. Indeed. This occurred with a hyperonymic pattern, rarely mentionned in existing studies. The pattern in question

consists of a definite or demonstrative determiner referring to a hyponym that appears earlier in the text. These cases are described in researches on reference (mostly not in the perspective of their use as patterns, which goes to show the importance for corpus linguists to take account of studies with an introspective approach), for example (Milner 1976) and (Lerat 1981). Such a relation is found in the following real example (18) (EDF corpus):

> *(18) Préparation des activités de liaison.*
>
> *Ce processus débute par la fourniture d'une copie de l'Etat de configuration du logiciel (processus is the hyperonym of préparation des activités de liaison)*
>
> *[Preparation for delivery. This process starts with the supply of a copy of the Status of the software configuration (process is the hyperonym of Preparation for delivery)].*

This pattern is difficult to describe and can hardly be applied in automatic analysis since the notion of "lack" plays an important part in it. It may be described as follows: [Defined or demonstrative determiner + N], under the following conditions:

- N does not appear in preceding context,
- the determiner does not have a deictic function,
- the definite determiner does not precede a noun with unique meaning, supposed known by the two interlocutors.

As for the the third aspect, the most difficult to take into account, we were only interested in examples with demonstratives. We proceeded in the following manner:

- Deleting the predictable deictic, referring to the writing process or to work in progress (*this book, this chapter, this part …*) ;
- Deleting prepositional phrases that we consider as set (*at this moment*, *with this aim*, etc. [1])
- Identifying sentences with a demonstrative followed by a noun under the condition that this noun does not appear in the preceding paragraph.

The results have not yet been thoroughly examined. Nevertheless, those we have already obtained are quite eloquent.

Tests have been performed on two corpora: a corpus supplied by the EDF, on the one hand, and a set of papers about knowledge engineering, on the other. In the

EDF corpus, there are 78 cases did not contain a demonstrative with a deictic function or a demonstrative within a set of prepositional phrase. The distribution is as follows:

- 44 examples contain a hyperonymic relation (between *activité* and *archivage de l'état de configuration* in (19)).

  (19) Archivage de l'Etat de Configuration logiciel. *Cette* activité *est à la charge du responsable de la gestion de configuration*

  [Archiving the software configuration status. The configuration manager carries out this activity.]

- 27 consist of a repetition of a predicate in noun form:

  (20)*Ce chapitre s'attache à* décrire *les différents processus de gestion de configuration. Cette* description *ne présume pas de la méthodologie de développement utilisée*.

  [This chapter sets out to describe the various configuration management processes. This description does not presuppose the development methodology used.]

- 4 consist of a repetition of an expression in noun form,

  (21)*Il se présente sous la forme d'un arbre dont les feuilles correspondent à des produits ou composants à réaliser. Cette* décomposition*, qui suit l'arborescence des produits définie au paragraphe 4.1…*

  [It is presented in the form of a tree whose leaves correspond to products or components to be carried out. This decomposition, which follows the tress structure shown in section 4.1 etc.]

- 2 corresponds to a synonymic relation,

  (22)*Le produit développé par l'équipe projet peut s'appuyer sur un* noyau logiciel *développé par une autre équipe projet. Cette dernière livre alors toutes les Unités de Configuration permettant à l'équipe projet de modifier cette* souche logiciel *en fonction des nouvelles fonctionnalités à implémenter.*

  [The project developed by the project team may be based on a software kernel developed by another project team. In this case, this second team makes

available all the necessary Configuraed Units to the first team to be able to modify this software stump depending on the new fonctions to be developed.]

- 1 example consists in a repetition of a noun phrase in acronym form.

  (23)*Le* Plan de Développement *standard définit les mécanismes de vérification et de validation de ces produits… Liens de traçabilité entre ce* PDL *standard et les propositions du MAQL-ST DER.*

  [The standard development plan defines mechanisms for verifying and validating these products. Traceable links between this standard PDL and the MAQL-ST DER proposals.].

Most of possible cases described in (Lerat 1981) are found but what is surprising is the frequency of hyperonymy: more than 56 %. So we can confer to this structure the role of pattern of hyperonymy anyway in this corpus and probably in others. Nevertheless, the same difficulty than with `chez` appears*:* relation is not posed, explained by the pattern, it is presupposed by it.

This study (still in progress) shows another kind of teaching concerning the link between pattern and aim of analyze. Actually, hyperonyms appearing in this structure are not at all the same than those appearing for example in a definition structure:

- First observation, in contrast with terms, generally considered as containing more than one word (in 70 to 80 % of cases), hyperonyms found with this pattern are simple terms in their majority. In EDF corpus, there are 12 polyterms on 44 hyperonyms (27 %). In Knowledge Engineering corpus, only 74 of 376 hyperonyms are polyterms (19,7%).
- Second observation: terms found in the structure with reference back are not of the same kind than the ones found in definition structure, neither in position of definiens (hyponym) nor definiendum (hyperonym), what is more surprising. In EDF corpus, no term found in the structure with a demonstrative is founded in position N1 or N2 in classical definition structure: [N1 être dét N2]. On the contrary, 8 of these terms are founded as head of one or more term(s) in position N1 or N2:

Acteur (actor) (1 term),

Activité (activity) (2 terms),

Composant (component) (2 terms),

Décomposition (decomposition) (2 terms),

Espace (space) (3 terms),

Phase (phasis) (2 terms),

Processus (process) (3 terms),

Revue (review) (1 term).

A hypothesis, which is still under verification, is that hyperonyms found in the structure with a demonstrative correspond to a higher level than those used in definition contexts. In some way, they are more domain-independent and closer to common language. In this way, they may be linked to the application since some applications need high-level terms, for example, indexing.

So, it seems that relation patterns may also be dependent on the nature of the application.

7- Conclusion

The aim of this paper was to highlight the notion of conceptual relation pattern with corpus analysis experiments. Three main conclusions may be retained from these experiments. The first one concerns the role of nature of the corpus, the second one concerns the role of application and the third one concerns pattern functioning.

1- Corpus role

As shown above, three kinds of dependencies may exist between conceptual relation patterns and corpus nature. Firstly, dependencies may be total. Conceptual relation patterns may be identified for a corpus but only for that corpus alone. In this case, from a linguistic point of view, it is necessary to build methods for identifying patterns (cf Condamines and Rebeyrolle, 2000). Secondly, dependency may be established between patterns and corpus extra-linguistic characteristics, in other words genre. In this case, a veritable corpus linguistic may be developed in order to identify regular expressions playing the role of patterns in corpora with specific characteristics. The aim is then to use systematically (and even automatically) these patterns in order to spot important relations. Thirdly, dependency may be almost non-existent if a pattern appears in all corpora, independently of the genre. This seems very rare and more

detailed studies should be carried out on varied corpora in order to prove this independence.

2- Application role

The last paragraph above describes an hyperonymic pattern used to identify very high level terms, or at the least, terms of different nature to those identified by other more traditional patterns. This observation opens reflections on the role of application on patterns. This role is well known for choosing relations but it would be necessary to examine it to describe patterns.

3- Pattern functioning

This paper describes several cases where relation is not expressed but presupposed. However, an element in the context is clearly the sign of the possibility to pass from discourse to modeling. From this observation I conclude that conceptual relation patterns have two roles. The first one consists in indicating the possibility of a transition from discourse to modeling. This possibility may be linked to corpus genre. The second role consists of helping relational interpretation. In this case, pattern meaning may be considered as bringing relation notion, on its own. This meaning may also be linked with corpus genre. In introspective studies, only this second role is indicated since it is the most obvious, but it is necessary to perform studies on the other role since it is as important as this one for identifying relations.

Numerous corpora, with different genres, are now available and it seems urgent to take them into account for all linguistic analysis. With examples on identifying relations, I hope I have showed here how such an approach can be fruitful and allow us to ask questions in an original way.

Note

[1] Note that in these phrases, the noun still plays the role of hyperonym and it was sometimes difficult to distinguish set expressions from others.

Bibiography

Ahmad, K. and H. Fulford. 1992. *Knowledge Processing : Semantic Relations and their Use in Elaborating Terminology*. Computing Sciences report CS-92-Guildford : University of Surrey.

Basili, R. and M.T. Pazienza. 1997. « Lexical Acquisition and Information Extraction ». In M.T. Pazienza (ed) : *Information extraction*. (Lectures Notes in Computer Science) Berlin : Springer-verlag, 44-72.

Benveniste, E. 1966. *Problèmes de linguistique générale 1*. Paris : Gallimard.

Biber, D. 1988. *Variation across speech and writing*. Cambridge University Press.

Bowden, P.R., Halstead, P. and Rose, T.G. 1996 « Extracting Conceptual Knowledge From Text Using Explicit Relation Markers ». In Proceedings of the European Knowledge Engineering Workshop (EKAW-96), *Lectures notes in Artificial Intelligence*, n°1076, Berlin : Springer Verlag. 146-162.

Condamines, A. 2000. « Chez dans un corpus de sciences naturelles : un marqueur de méronymie ? ». *Cahiers de Lexicologie* n° 77, 2000-2, 165-187.

Condamines. A. and P. Amsili. 1993. "Terminology between Language and Knowledge: an example of Terminological Knowledge Base", TKE 93 (Terminology and Knowledge Engineering), Cologne, août 1993. Xx-xx

Condamines, A. and J.Rebeyrolle. 2000 " Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode ". In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, (eds). *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles. 127-147.

Condamines, A. and J.Rebeyrolle. : " Searching for and Identifying Conceptual Relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results ". D.BOURIGAULT, M.C. L'HOMME, C.JACQUEMIN (eds) : *Recent Advances in Computational Terminology*, Amsterdam/philadelphia :John Benjamins. Xx-xx

Cruse, D.A. 1986. *Lexical Semantics* ; Cambridge : Cambridge University Press .

Deville, G. 1989. *Modelization of Task-oriented Utterances in a Man-Machine Dialogue System* . Thèse de Philologie, Université d'Antwerpen.

Gaizauskas, R., and Y. Wilks. 1988. « Information Extraction : Beyond Document Retrieval » . *Journal of Documentation*, vol.54, n°1. 70-105.

Lerat, P. 1981. « Les noms de relation », *Cahiers de lexicologie*, 39-2. 55-65.

Lyons, J. 1978. *Eléments de sémantique*, traduction de J.Durand. Paris : Larousse.

Martin, R. 1983. *Pour une logique du sens*. Paris : PUF.

Meyer, I., Douglas, S., Bowker L., and K. Eck. 1992. “Towards a new generation of terminological resources: An experiment in building a terminological knowledge base”. In Proceedings 16th *International Conference on Computational Linguistics*. Nantes. 956-957.

Meyer, I. 2000. Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework. D.BOURIGAULT, M.C. L’HOMME, C.JACQUEMIN (eds) : *Recent Advances in Computational Terminology*, John Benjamins. 279-302.

Milner, J.C. 1976. Réflexions sur la référence. *Langue française* n°30. 63-73.

Pearson, J. 1998. *Terms in Context,* Amsterdam and Philadelphia: John Benjamins.

Rebeyrolle, J. 2000. *Forme et fonction de la définition en discours* ; Thèse de Sciences du langage, Université Toulouse Le Mirail.

Riloff, E. 1996. « Automaticaly Generating Extraction Patterns from Untagged Text », in Proceedings of the *Thirteenth National Conference on Artificial Intelligence* (AAAI-96). 1044-1049.

Séguéla, P. 2001. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques* ; Thèse d’informatique, Université Paul Sabatier, Toulouse.